

European Language Resources Association

History and Recent developments

Dr. Khalid CHOUKRI

ELRA/ELDA

55-57 Rue Brillat-Savarin, 75013 Paris, France

Tel. +33 1 43 13 33 33 – Fax. +33 1 43 13 33 30

Email: choukri@elda.fr

Web: <http://www.elda.fr> or <http://www.elra.info>

1	INTRODUCTION.....	2
2	EUROPE FROM 1951 TREATY TO THE EUROPEAN UNION.....	3
3	EUROPEAN R&D FRAMEWORKS.....	4
3.1	GENERAL STRUCTURES.....	4
3.2	HUMAN LANGUAGE TECHNOLOGIES / LANGUAGE ENGINEERING RELATED ACTIONS.....	5
3.3	CAPITALIZATION ON THE EU PROGRAMS.....	6
4	ELRA FOUNDATION.....	6
4.1	RATIONAL BEHIND ELRA.....	6
4.2	MEMBERSHIP DRIVE	7
5	ELRA’S MISSION AND TECHNICAL ACTIVITIES.....	7
5.1	IDENTIFICATION OF USEFUL RESOURCES.....	8
5.1.1	<i>Description of Language Resources.....</i>	<i>8</i>
5.1.2	<i>ELRA Language Resources Catalogue.....</i>	<i>8</i>
5.1.3	<i>The Basic Language Resources KIT (BLARK).....</i>	<i>10</i>
5.2	HANDLING THE LEGAL ISSUES RELATED TO THE AVAILABILITY OF LANGUAGE RESOURCES.....	11
5.3	DISTRIBUTION ACTIVITIES AND PRICING POLICY.....	12
5.4	VALIDATION AND QUALITY ASSESSMENT.....	14
5.5	COMMISSIONING THE PRODUCTION OF NEEDED LANGUAGE RESOURCES & MARKET WATCH.....	15
5.6	HUMAN LANGUAGE TECHNOLOGIES EVALUATIONS.....	15
5.6.1	<i>Carrying out Evaluation Campaigns and/or Providing Evaluation packages for all Human Language Technologies.....</i>	<i>15</i>
5.6.2A	<i>European Evaluation Infrastructure within ELRA.....</i>	<i>16</i>
5.7	INFORMATION DISSEMINATION, PROMOTION AND AWARENESS:	17
5.7.1	<i>Major Conferences established by ELRA: LREC & LangTech.....</i>	<i>17</i>
5.7.2	<i>Euromap Language Technologies.....</i>	<i>17</i>
5.7.3	<i>NEMLAR: A Network for Euro-Mediterranean Language Resource and human language technology development and support.....</i>	<i>19</i>
6	ELRA COLLABORATIONS AND PARTNERSHIP.....	19
7	CONCLUSION.....	20

1 INTRODUCTION

This paper aims at describing the rationale behind the foundation of the European Language Resources Association (ELRA) in 1995 and its activities since then. Part of this paper is an update of an invited talk given at the 1999 Oriental COCODA WORKSHOP (more information about the workshop at: <http://www.milab.is.tsukuba.ac.jp/o-cocoda/cocoda99/content.html>)

In order to better understand the European landscape, this paper briefly elaborates on the recent history of the European Union and its R&D programs of interest to the Human Language Technologies Community.

After this, the paper focuses on the issues involved for making language resources available to different sectors of the language engineering community as well as contributing to the evaluation of Human Language Technologies (HLT). ELRA is presented as a conduit for the distribution of speech, written and terminology databases, enabling key players to have access to Language Resources (LRs) for technology development and technology evaluation. In order to effectively produce and provide such resources to research and development groups in academic, commercial and industrial environments, it is necessary to address legal, logistic and other practical issues. Therefore ELRA's initial mission was to set up a centralized Not-for-profit organization for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role of a central repository, ELRA had to address issues of various natures such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, Intellectual Property Rights (IPR), Industrial rights, etc.), information dissemination (to act as a clearing house). ELRA has set up an operational body (ELDA: Evaluations and Language Resources Distribution Agency) to take care of the daily aspects related to these missions.

Since its establishment in 1995, ELRA has managed to make available, worldwide, a large set of marketable resources. ELRA handled the legal issues through generic license agreements and IPR manuals that were made widely available. A set of validation manuals have been produced and are widely distributed (in particular for speech and written resources), to promote quality and best practices.

After the first phase, ELRA has been actively promoting the evaluation of Human Language Technologies through the establishment, via ELDA of an evaluation infrastructure in Europe. Drawing on its experience in national and Europe-wide evaluation projects and also its experience in the production, validation, packaging and distribution of language resources, ELDA's evaluation department is working to strengthen a European clearing house for evaluation related resources and software packages, in the same way that ELDA has become the European clearing house for language resources. ELDA's vision for a European evaluation infrastructure is inspired by both European and international evaluation initiatives, including the DARPA/NIST evaluation programme in the United States.

The association now has a good number of members from all over the world (including Asia) who benefit of a number of services in addition to substantial discounts on the resources presented in the catalogue. Each member belongs to one of three possible colleges (speech, written or terminology) in accordance with its principal area of interest. Legal and contractual assistance are also included in the membership as ELRA managed and continues to work to simplify the relationship between providers and users (customers) by drafting generic agreements that specify the responsibilities and duties of each party when licensing a language resource. More than 125 agreements with providers of language resources have been secured by ELRA, preventing users and providers from spending their time on contractual negotiations. In order to add value to the resources it distributes, ELRA also initiated the production of validation manuals for each resource type, namely spoken resources, written corpora, lexica and terminology resources. Both on-line and hardcopy versions of the validation manuals are available; LR providers are encouraged to use these manuals to specify or validate their databases. In order to efficiently carry out such work, ELRA established a network of validation units among the most skilled laboratories in Europe, involved in Speech and Written Resources.

An important highlight of the association work on promoting the Human Language Technologies area is the set up of the International Conference on Language Resources and Evaluation (LREC). This conference was initiated by ELRA and organized in co-operation with other national and international associations and consortia. The first Edition took place in Granada, Spain, with over 500 attendees, and was a major success when one considers that it addressed the specific topic of language resources and evaluation. The second edition took place in Athens, Greece (2000) while the third one took place in Las Palmas, Spain (2002) with over 700 participants.

In addition to this, ELRA has strongly contributed to the set up and the organization of LangTech (Berlin, 2002 and Paris, 2003) which is the European Forum for Speech and Language Technology. LangTech is organized to promote the business behind the HLT. With more than 350 industrial and academic representatives and over 21 exhibiting companies at LangTech2003, LangTech is proving to be a successful and fruitful forum. LangTech 2003 brought together some of the major European and international companies, with a programme featuring over 50 presentations given by industrial leaders and academic key players, the participation of 12 start-ups in the "elevator pitch" session and an exhibition hall where 21 companies working in various areas of Natural Language Processing and Knowledge Management were able to meet their actual and potential customers.

As for the validity of the existence of ELRA for the collection and the distribution of language resources, the sales statistics from the recent years speak for themselves:

- *A large number of agreements were signed with providers of language resources. By end of 2002, ELRA's catalogue of language resources consisted of over 725 items of Spoken Language Resources, Written Language Resources (mono- and multilingual lexicons, corpora), and terminology databases. Worth to be mentioned is also the fact that some of these resources (SLR or WLR) cover languages that were not yet represented in our catalogue, e.g. Korean, Basque or Turkish.*
- *For the distribution of LRs, in 2002, 481 LRs were sold, compared to 406 in 2001 and 266 in 2000. The 2002 fiscal report showed a significant growth. Out of the total of 481 sales for the year, 347 were for research and 134 were for commercial purposes. Of these 481 purchases, 334 were specifically speech databases, including about 135 Aurora packages for Speech front-end evaluation.*

ELRA is now in the position to establish close partnership and Joint-ventures, in particular with similar organization in Asia. It is of paramount importance that ELRA reaches fruitful agreements with other regional organizations in order to achieve, all together, a better streamlining of efforts in the development of new Language Resources that are of interest to "global" players.

2 EUROPE FROM 1951 TREATY TO THE EUROPEAN UNION

The idea of a united Europe emerged after the 2nd World War as a barrier against future wars. A first attempt to establish a customs union between France and Italy failed and led the French foreign minister Robert Schuman to think of a less ambitious structure, restricted to a small number of commodities linked to basic industries. This was thought of as an essential first step towards the Union. As the idea that coal and steel were the key to economic development and that international security implied that military dangers would be reduced if such crucial industries were under international control, Jean Monnet conceived the European Coal and Steel Community (ECSC) in 1950, and it was established in April 1951 by the treaty of Paris, signed by six countries: France, Germany, Italy, Belgium, The Netherlands, and Luxembourg. From the beginning it was designed as a supra-national organization rather than an inter-governmental committee, with high level and independent authorities. The first organization provided a model for the European Economic Community and its success encouraged European leaders to extend the model to other areas such as energy and agriculture. In 1957, the treaty of Rome established the European Economic Community (EEC) through the foundation of common institutions with the idea of a progressive fusion of national institutions, a harmonization of all policies and an evolution towards a common market. The EEC lived in parallel with the 1951 treaty and the ECSC as well as with the Atomic union (EURATOM), including some duplicate structures. Ten years later, in April 1967, a new treaty (the so-called the Merger Treaty) led to the establishment of a single framework referred to as the European Community (EC) with a large set of common policies.

In 1973, three new countries joined the EC: Denmark, Ireland and United Kingdom. Greece joined in 1981, Spain and Portugal in 1986, Austria, Finland, and Sweden on the 1st of January 1995, and the EC thus grew to a total of 15 countries. Several other countries have made their formal application to join the Union, in particular countries from the Eastern and Central Europe.

In 1985, the decision to create a common market was agreed upon with a deadline to remove inter-European borders by 1992. This creates an area without any borders in which goods, services, capitals, and people can freely circulate, with a combined population of over 370 million inhabitants.

In 1992, the 15 countries agreed on a new treaty (the so-called Maastricht Treaty) which should implement "a closer union among the peoples of Europe" with strong and reinforced responsibilities of supra-national organizations (e.g. the Executive body called the European Commission and the European parliament). The treaty was adopted by all the countries and went into effect on the 1st November 1993. After this treaty, the European Community (EC) became the European Union (EU). The major clause of the Maastricht Treaty is the establishment of a time table and a set of criteria for the introduction of a single currency: the EURO (€) and a European central bank to coordinate economic convergence and to promote monetary cooperation, a very challenging task, successfully carried out and achieved by the 1st of January 1999 for 11 countries (all except UK, Sweden, Denmark, and Greece). A transition period was agreed until 1st July 2002, when the EURO will become the sole legal tender in Europe. During that period, exchange rates between the EURO and the other national currencies are irrevocably fixed. EURO paper currency and coins have been issued on the 1st January 2002 and replaces all national ones on the 1st July 2002. The Maastricht Treaty also included a European citizenship with voting rights to local and European parliament elections. A new treaty (The Amsterdam Treaty, entry into force: 1st May 1999) contains new rights for citizens, freedom of movement, employment, strengthening the institutions, etc.

The last treaty is the one known as the Nice Treaty which amended the Treaty on European Union to consider the accession of 10 new countries to the Union by May 2004, bringing the European Union to 25 countries. Nice Treaty will enter into force on 1 February 2003 and will govern the enlargement of the European Union.

The European high-level institutions are the European parliament, The Council of the European Union, The European Commission, and The Court of Justice. Basically, since 1979, the Parliament has become directly-elected and acquired new powers. Instead of the advisory role it has played for its first twenty years, its role is now "its primary objectives are like the ones of any Parliament - to pass good laws and to scrutinize and control the use of executive power. Now

more than ever before, it is in a much better position to do both because its responsibilities have been gradually widened and its powers strengthened first by the Single Act of 1987 and then by the Treaty of European Union of 1993, to include legislative power and budgetary powers". More official information is available at: <http://europa.eu.int>.

The Council of the European Union (usually known as the Council of Ministers) is a body with the characteristics of both a supranational and intergovernmental organization. Here, the Member States legislate for the Union, set its political objectives, coordinate their national policies and resolve differences between themselves and with other institutions. The council consists of one representative from each country – in principle the Minister of foreign affairs or the Minister in charge of the matters under discussion. The presidency of the council rotates between the Member States every six months.

The European Commission (also referred to as the Commission) identifies three distinct functions: initiating proposals for legislation, guardian of the Treaties, and the manager and executor of Union policies and of international trade relationships.

The Court of Justice provides the judicial safeguards necessary to ensure that the law is observed in the interpretation and application of the Treaties and, generally in all of the activities of the Union.

The Research and technological development programs, vital for the future of Europe, are orchestrated by the Commission. The Commission is divided into a number of Directorates-General (DGs) with additional specialized services. Each DG is headed by a Director-General, reporting to a Commissioner who has the political and operational responsibility for the work of the DG. Some of the DGs that contribute to the Language Engineering are: The Information Society DG, Research DG and the Education and Culture Directorate-General.

For instance the Information Society Directorate-General is playing a key role in implementing the 'vision' set by Europe's heads of state in Lisbon, 2000: "to make Europe the world's most competitive and dynamic economy, characterised by sustainable growth, more and better jobs and greater social cohesion, by 2010". The Information Society Directorate-General stimulates research into Information Society technologies which can be integrated into the citizen's everyday environment, business and administration. Many issues related to HLTs are part of the mission of this DG.

The mission of the Research Directorate General's is to develop the European Union's policy in the field of research and technological development and thereby contribute to the international competitiveness of European industry; It is the DG that coordinate European research activities and conduct the research & technological development at the European Level. One of the instruments used for the implementation of this policy is "the multi-annual Framework Programme which helps to organise and financially support cooperation between universities, research centres and industries". The current Sixth Framework Programme covers the period 2002-2006 and has a total budget of €17.5 billion, with about €3.6 billion to be devoted to Information Society Technologies.

Another directorate of interest to us here is the Education and Culture Directorate-General's which has as main mission the building of a " Europe of knowledge" and "preserving and enhancing Europe's cultural diversity in the various fields where it is expressed, in particular through measures to support the competitiveness of the European audiovisual industry and promote **linguistic diversity and language learning**".

For more information about the Commission role and structure, including possible partnership with third countries (in particular the External Relations Directorate General which is in charge of the Central Asian Republics; programming Tacis assistance to Eastern Europe, Caucasus, Central Asian Republics; and relations with Asia), please see: http://europa.eu.int/comm/index_en.htm

3 EUROPEAN R&D FRAMEWORKS

3.1 General structures

As said above, the European Commission is in charge of the research and technological development programs. The R&D activities were listed on the common activities of the Community from the beginning. The 1951 ECSE Treaty encouraged technical and scientific research in the steel and iron industry. The 1957 Euratom Treaty established the Joint Research Centre (JRC), the cost-sharing contract research program and procedures for the coordination of national research projects. In 1982, the Council of the European Union affirmed the need to "systemize and optimize Community activities in the field of research, development and demonstrations by the adoption of a framework program containing broad indications for the medium-term development of the Community's scientific and technical objectives". In 1983, the Council approved the principle of framework programs for periods of **four years** and defined the scientific and technical objectives and selection criteria for the period 1984-1987. Each program is split into a set of subdivisions ranging from developing agricultural productivity to New Technologies and other actions related to improving living and working conditions, etc. The idea is to fund projects submitted by consortia of European organizations on the basis of shared costs between the organizations and the European Commission.

In the meantime (1983), a pre-framework was launched as a preparatory action in the field of Information Technologies (IT) by DGIII. The program acronym was ESPRIT 0 and referred to European Strategic Program for Research and development in

Information Technologies. ESPRIT programs focus on the Electronics, Microelectronics, Information processing and Information systems.

The ESPRIT 0 was of 12 months duration and was allocated a budget of 11,5 Million ECU (European Currency Unit which became now the EURO, which is approximately 1.20 US\$). ESPRIT 1 to 4 were parts of the framework programs 1 to 4 even if other actions were also addressing topics of interest for Language Engineering. The ESPRIT 1 budget was 750 Million ECU and lasted 60 months (1984-1988) with about 241 projects. The ESPRIT 2 budget was 1600 Million ECU and also lasted 60 months (1987-1992) with about 435 projects. The ESPRIT 3 budget was 1532 Million ECU and lasted 42 months (1991 - 1994) with about 605 projects. ESPRIT 4 started in 1994 for a period of 49 months with a budget of 2084 Million ECU and granted funding for about 1341 projects (for more details: <http://www.cordis.lu/esprit/src/intro.htm>).

The 1st Framework Program (FP1), lasted 48 months (01/01/84-12/31/87) and had a budget of 3750 million ECU of which 680 Million ECU was devoted to new Information technologies. It was managed by DGXIII. The second framework program (1987-1991) had a budget of 5396 million ECU with about 2275 Million ECU devoted to information and communication society items (IT, Telecommunications, new services). The third framework program (1990-1994) had a budget of 6600 million ECU with a large share devoted to Information and Communication Technologies ICT (2516 Million ECU).

The fourth program had a budget of 11879 Million of ECU and covered the period 1994-1998. Some of the projects initiated during that period are still running. The portion devoted to Information and Communication Technologies is 3626 Million ECU with a large part devoted to Telematics, ACTS (Advanced Communications Technologies and Services 671.00 million ECU), and ESPRIT 4.

The Fifth framework program had a total budget of 14960 Million EURO of which 3925 Million EURO are devoted to the action called: Creating a user-friendly information society (known as **IST**). This program covers the period 1999-2002.

The current Framework program is the 6th. As for the other FP programmes, the 6th Framework Programme "has to serve two main strategic objectives: Strengthening the scientific and technological bases of industry and encourage its international competitiveness while promoting research activities in support of other EU policies. The 6th FP budget is about **17.5 billion euros for the years 2002 - 2006** among which 3.6 Billions will be allocated to the IST. Here we do not elaborate on national initiatives in R&D (some of the current HLT programmes at national levels will be pointed out in the coming sections) but one should bear in mind that the FP6 budget "represents about 4 to 5 percent of the overall expenditure on RTD in EU Member States". Therefore the commission has decided to devote €270 Millions to stimulate the coherent development of research and technology policy in Europe by supporting programme co-ordination and joint actions conducted at national and regional level as well as among European organisations.

3.2 Human Language Technologies / Language engineering related actions

Specific action lines addressing Language Engineering issues have been funded within all the framework programs. The initial ones were funded under the ESPRIT framework.

A well known action within the second framework program is the EUROTRA project which aimed at creating an advanced machine translation system. This project intended to deal with all official languages of the Community by producing an operational system prototype in a limited field and for limited categories of text, which would provide the basis for subsequent system development on an industrial scale. The project lasted about 10 years (82-92) with a budget of 37 Million ECU.

Within the third framework program we had another action line called LRE which focused on the development of linguistic resources and related computational tools (Software tools; Grammars; Dictionaries; Terminological collections; Text corpora), the support for the formulation of standards and guidelines for the encoding and interchange of linguistic data), pilot and demonstration projects (Machine translation, Document abstracting and indexing; Aids for mono- and multilingual document generation, storage and retrieval; Man- machine communication; Construction of knowledge bases from natural language text; Computer-aided instruction), etc. The official objective was to develop basic linguistic technologies which would be incorporated by European industrials into a large number of computer applications where natural language is an essential ingredient. This action started in 1991 and lasted until 1994 with a budget of 22.50 million ECU.

Within the fourth framework program, a new action called Language Engineering was launched and focused on the integration of new oral and written language-processing methods and general linguistic research that aimed at improving possibilities for communicating in European languages.

The largest specific program in the FP5 was the IST program with a budget of 3600 Million EURO. This program aimed at bringing together all previous Community ICT activities: ESPRIT, ACTS and Telematics. The IST Program was structured as four inter-related key actions. The Key Action III (KA III) on Multimedia Contents and Tools includes a specific activity called HLT: Human Language Technologies which focuses on the usability and accessibility of digital content and services while supporting linguistic diversity in Europe. Key action III had a budget of 564 Million over the four years.

Within the current FP6, there is a focus on Knowledge and interface technologies with three main hot topics: Multimodal interfaces, Semantic-based knowledge systems and Cognitive systems.

3.3 Capitalization on the EU programs

Several reports have been written about the breakthroughs allowed by the European Community R&D programs. The list of projects and programs is available at the Cordis URL: <http://www.cordis.lu/>. A large number of today's products and/or advanced prototypes were developed in the frameworks mentioned above. The European Commission had a pre-requisite for all the proposals to include a section about commercial and industrial exploitation of the project foreground. This foreground consisted of all outcomes of the project: system components, software packages, hardware platforms, best practices, and language resources collected within the projects as one of its requirements.

Due to the funding scheme for projects (shared costs, funding from the commission not to exceed 50% of the total budget, projects' duration of 2 to 5 years), several costly resources were collected but were never disseminated. In a large number of cases the data was completely lost.

Some of us may remember very famous projects such as SAM which aimed at providing a pan-European basis for the assessment of speech technology devices and led to the development of speech databases for almost all Western European languages (the so called EUROM 1 databases). SAM was funded under ESPRIT 1 and ESPRIT 2. Some of the databases produced within SAM are still not available today. Another project is SUNDIAL funded in the framework of ESPRIT 2. SUNDIAL addressed the problem of speech-based cooperative dialogue as an interface for computer-based information services. The main technologies that were to be developed were continuous speech recognition and understanding, and oral dialogue modelling and management. SUNDIAL also produced several speech databases which are not accessible today. The project ended in 1993. Another project that concerned speech is POLYGLOT which aimed at the development of multilingual voice interface (input/output) for a number of commercially promising applications for 6 European languages. The project has produced speech databases for speech to text and text to speech for 6 European languages. These are not available today although the project ended in 1992. We can quote many other projects.

Many organizations involved in these projects restructured their activities; others withdrew from the field, and after a couple of years the archives were cleaned up and the databases were completely lost: a commodity worth millions of EURO!! This was no more acceptable for the sponsors nor for the key players in this area.

4 ELRA FOUNDATION

4.1 Rational behind ELRA

The sponsoring agencies, in particular the DGXIII (now DG Information Society) and its Language Engineering sector, considered that a successful development of language technologies is and will continue to heavily rely on the availability of large scale Language Resources, together with appropriate standards, methodologies, and manipulation tools (see reference Eagles Spoken book). They also considered the need to capitalize on all the investments done on the production and the packaging of LRs in order to ensure certain re-usability. The access to large annotated spoken and written corpora, together with appropriate higher-level LRs, would not only provide a direct benefit to research and development efforts across a wide range of private and public organizations, but would also foster fruitful academic and industrial cooperation.

In order to carry out such strategic analysis, the European Commission and a large number of key players in the Language Engineering field decided to launch a project called RELATOR. The project aimed at defining a broad organizational framework for the creation of language resources, for both written and spoken language engineering, which are necessary for the development of an adequate language technology and industry in Europe. It also aimed at determining the feasibility of creating a co-ordinated European network of repositories which would perform the function of storing, disseminating and maintaining such resources. This activity is intended to contribute towards the long-term goal of making large scale LRs widely available to European organizations involved in R&D and educational activities.

The major outcome of RELATOR was the creation of ELRA as well as the initiation of several Language Resource production projects (e.g. SpeechDat family, PAROLE, etc.). This action was carried out through discussions in which high level representatives of the relevant actors were involved (e.g. owners of resources, producers, private and public users, funding bodies, scientific and professional associations). The big industrials as well as the major R&D labs were represented in what was known as the RELATOR Steering Committee, chaired by the Deputy Director General of the DGXIII. The discussions were about the various aspects of the problem, their needs and requirements, the possible solutions, their willingness to co-operate, and the conditions for a joint European action. They also aimed at identifying, describing and evaluating at various levels (e.g. organizational, technical, legal) the alternative methods and structures which could ensure the creation, management and maintenance of a European repository of re-usable LRs, and their dissemination to the various types of users.

The project had to present final recommendations for establishing a collaborative infrastructure that will act as a collection, verification, management and dissemination centre, built on the foundation provided by existing European structures and organizations.

The project consortium comprises representatives of major European-wide bodies and associations, most notably ELSNET, ISCA and EACL, plus an industrial steering committee composed of representatives of leading IT companies, publishers, Telecom

operators and other providers of electronic information services. The action has been carried out in co-operation with relevant European groups and with on-going initiatives such as EAGLES. It also implied an analysis of existing international structures. An unanimous decision was made to set up an independent, permanent, centralized organization. Legal experts were asked to evaluate different scenarios and to suggest the most useful one. Of all suggestions (an incorporated company, a grouping of economic interests as allowed by the French law, a non-profit association) the steering committee felt that an independent, not-for-profit, membership-driven association was the easiest to implement and the most acceptable from the liability and the legal points of view. The association statutes were prepared and deposited in Luxembourg on February 1995 by 16 founding members, under the name of **The European Language Resources Association (ELRA)**.

As an association, ELRA is governed by a Board, consisting of twelve members, elected by an open vote of all ELRA members. The Board defines the association's goals. The policy is implemented by the Chief Executive Officer (CEO) and his staff.

In order to efficiently carry out the distribution activities, the CEO has set up the Distribution Unit called Evaluations & Language resources - Distribution Agency (ELDA) as the organizational infrastructure for identifying, classifying, collecting, validating, and distributing LRs. ELDA is responsible for the development and execution of the ELRA strategies and plans as defined by the Board. It is incorporated as a company in order to handle all the commercial and business oriented tasks. In particular, ELDA handles the legal matters that arise in connection with the distribution of LRs and provides legal advice to ELRA members and suppliers on these topics. Agreements for distribution purposes are drafted and entered into on behalf of ELRA. ELDA has also launched and contributed to a number of evaluation campaigns in various areas of HLTs.

ELRA also seeks to harness the expertise of its members via a number of informal panels of experts (e.g. panel of experts in charge of the identification of existing resources or the panel of experts in charge of the validation of language resources). The panels advise the Board and the CEO in specific areas and consist of a core of ELRA members.

4.2 Membership drive

At the beginning only institutions legally registered in Europe were able to join ELRA. Since 1997 this restriction has been lifted and any legally registered organization can join the association. Full membership, with voting rights, is available only to organizations legally established in Europe.

During the General Assembly of 28 November 1997, the members agreed to the proposal of the Board to modify the annual membership fee as follows:

Non-profit making organizations	750 EURO
European small/medium-sized companies < 50 employees	1000 EURO
European profit making organizations >= 50 employees	1500 EURO
Non-European profit making organizations	5000 EURO

Purely for organizational purposes, members are classified by their main interest (spoken, written, or terminological resources). The difference between membership fees of European and Non-European companies reflects the grants received from the European agencies.

Since its foundation, ELRA has attracted an important and steady number of members as shown below:

	1995	96	97	98	99	2000	01	02
Paid-up members, inc.								
free membership	63	70	66	71	79	95	93	101

The services offered by ELRA to its members are summarized both on the ELRA web site and brochures. These services go beyond the important discount given on the price of language resources.

5 ELRA'S MISSION AND TECHNICAL ACTIVITIES

As stated above, ELRA has been entrusted with a crucial mission: to ensure that Language Resources needed by Language Engineering players are made available when they already exist or to produce them in a cost-effective frame. This mission is tuned from time to time to anticipate future requirements. Such a mission can be itemized as:

◆ Language Resources related issues

- The identification of useful resources.
- Handling the legal issues related to the availability of Language Resources.
- The distribution activities and Pricing policy.
- The validation and Quality assessment.

Commissioning the production of needed Language Resources & Market watch.

◆ Evaluations of Human Language Technologies

Commissioning the production of needed Language Resources for evaluations
Carrying out Evaluation Campaign
Providing Evaluation packages for all Human Language Technologies

◆ Promotion of the field, information dissemination and awareness, market watch and analysis

5.1 Identification of useful resources

5.1.1 Description of Language Resources

In order to play its role, ELRA committed to create structured and publicly available catalogues of Language Resources. In order to do so, ELRA has prepared a set of description forms to help the providers describe what they propose to ELRA for distribution in a more uniform and consistent way (see the URL corresponding to the catalogue at: <http://www.elda.fr/rubrique6.html> and the related forms at <http://www.elda.fr/leg/spee-en.ps> for speech resources).

This form includes all the features one need to know about a speech database: file standards, acquisition conditions (Telephone vs Microphone, environment), annotation levels, etc. Extensive work is being carried out worldwide on these "metadata" issues. The main projects are the OLAC (OLAC; the Open Language Archives Community), IMDI and more recently INTERA, a coordination project in which ELDA is playing an important role and which is funded by the European Commission. OLAC "is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources". (for details: <http://www.language-archives.org/>)

IMDI, which stands for "the EAGLES/ISLE Meta Data Initiative", aimed at making proposals for a standard of metadata descriptions of Multi-Media/Multi-Modal Language Resources. By now a number of metadata proposals have been released and using such a standard it is possible to create a browsable and searchable universe of such resources in the Internet (For more details: <http://www.mpi.nl/IMDI/>).

INTERA is building on the achievements of IMDI and targets the use of a customized IMDI metadata by most of the European centers involved in producing, sharing and/or distributing Language Resources (more details <http://www.mpi.nl/INTERA>).

5.1.2 ELRA Language Resources Catalogue

In the first ELRA catalogue, the identified resources were at different status levels: available through ELRA, available through the owner, under negotiation but not available yet, available on a case by case basis, identified but not available, etc. There were simply too many categories which made things too complex to be understandable.

Since then we committed to publishing a catalogue of resources that are available via ELRA or, in very few cases and for very sensitive databases, available through the owner/provider. Our catalogues are compiled with respect to the three colleges of ELRA: speech, written, and terminology. Some tools can be also catalogued if they are available for free. Very recently we have decided to add a fourth category to our catalogue to take into account new emerging trend of multimodal/multimedia resources.

The 4 categories are: "Speech and Related Resources", "Written Resources", "Terminological Resources", and "Multimodal/multimedia resources". They are also split into sub-categories as indicated below:

1/Spoken LRs

A - Telephone or Desktop/Microphone recordings

The databases catalogued in this section have been produced with speakers' recordings made over the telephone (fixed or mobile) network, or through a (set of) microphone(s). It consists of speech resources recorded in various environments, and covering a large number of European and non-European languages, e.g. the databases produced in the framework of the SpeechDat project, the BABEL databases, produced in the framework of the Copernicus programme, consisting of spoken corpora recorded in a controlled office environment, etc.

B - Speech Related Resources

This section consists of pronunciation and phonetic lexicons, the BDLEX, PHONOLEX, and MHATLEX databases, etc.

2/Written LRs

A - Corpora

This category contains monolingual and multilingual corpora, parallel or not, which may also be annotated. A few examples of the kind of resources one may find in this section are e.g. the corpora developed in the framework of the MULTTEXT project, the Multilingual and Parallel Corpora (MLCC), French scientific corpora, newspaper corpora in Arabic, etc.

B - Monolingual lexicons

The sub-category dedicated to monolingual lexicons contains various types of dictionaries, e.g. a dictionary of French verbs, some of the PAROLE lexicons in many languages, etc.

C - Multilingual lexicons

Here one can find either bilingual or multilingual dictionaries or lexicons, including e.g. the EuroWordNet database, etc.

3/Terminological LRs

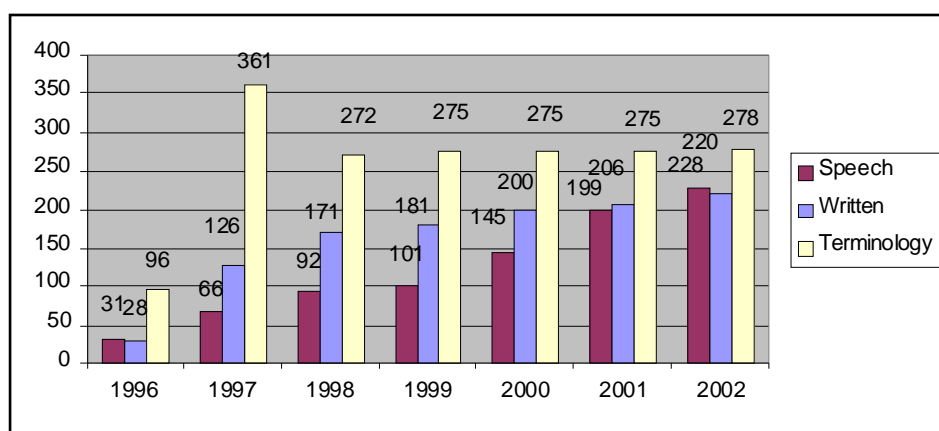
Several Monolingual, bilingual and multilingual terminological databases are available. They cover a large number of specialised domains, e.g. automobile engineering, insurance, linguistics, finance, etc., in a wide variety of languages.

4/Multimodal/Multimedia LRs

The resources of this new category have been produced using different modalities, including the speech. An example of such resources is the database produced in the framework of the M2VTS project.

In the under-completion revision of the catalogue we have decided to have a Language Resources Catalogue and a new catalogue of HLT Evaluation packages.

The progress of our identification task is illustrated through the following histogram indicating the number of items secured for distribution by the end of each calendar year:



(Some resources have been deleted from our catalogue due to our quality assessment process that rejected the databases as being of very low quality. Some other resources have been taken out due mainly to mergers of providers which led to the termination of some distribution contracts).

In the speech area, we see that the catalogue has grown from the 31 initial resources of 96 to more than 228 today. This should not hide the fact that many key resources are still not available for a large number of languages (including Western European ones). If we consider some basic resources such as:

- Articulatory database (e.g. ACCORD)
- Basic speech data with some phonetic material and some phonetic sequences, by a small number of speakers, recorded in a quiet environment (EUROM 1 & BABEL)
- Pronunciation lexicon (BDLEX)
- Proper names pronunciation lexicon (ONOMASTICA)
- Newspaper read text (BREF, Siemens-100, Apasci)
- Basic telephone speech (SPEECHDAT)
- Telephone-based speaker verification. (PolyVar)
- text corpora for language models (MLCC, Le Monde ...)

and if we consider some languages such as French, German, English (from several countries), an extract from the complete matrix is given below and illustrates that many basic resources (as defined by ELRA) are not available for distribution or do not exist at all.

Speech Resources	fre-fr	Fre-be	Fre-sz	fre-lu	Fre-ca	fre-int	eng-gb	eng-us	eng-int	ger-de	ger-at	ger-lu	ger-int	ita-it
Broadcast speech							E	e, t	E	E				E
Articulatory database	E				E		E			E				E
Microphone/desktop speech	E		E				E	e	E	E			E	E
Read newspaper texts	E									E				E
Telephone speech database	E	E	E	E	E		E	E	E	E		E		E
Mobile-radio speech								e						
Pronunciation lexicon	E							e		E				E
Onomasticon	E						E	e		E				e
Speaker identification speech corpus			E					e						E

Legend:

E : available through ELRA e : exists "blank": not identified/ does not exist t : transcribed

This matrix illustrates that many basic resources are not available and that there is a need to stimulate their production in order to meet the needs and requirements of both academic institutions and industrial users.

5.1.3 The Basic Language Resources KIT (BLARK)

Based on this, ELRA started to promote a concept of a Basic Language Resource Kit (BLARK) for all languages and later on extended this concept towards an Extended Language Resource Kit (ELARK) that may be useful for some languages that can afford to have more than the basics.

In the framework of the ENABLER thematic network (European National Activities for Basic Language Resources - Action Line: IST-2000-3.5.1), ELDA elaborated on a report defining a (minimal) set of LRs to be made available for as many languages as possible and map the actual gaps which should be filled in order to meet the needs of the HLT field. The report aimed at providing the basics on a larger initiative in order to determine the BLARK concept more specifically. The BLARK concept (Basic Language Resource Kit) was first launched in The Netherlands, through ELSNET (European Network of Excellence in Language and Speech). A Dutch initiative, called Dutch Human Language Technologies Platform was initiated in April 1999 by the Dutch Language Union (Nederlandse Taalunie), to stimulate collaboration between all actors involved and co-operation between Flanders and the Netherlands, and also at encouraging Flemish and Dutch participation in European projects and initiatives.

Further to their own initiative, without mentioning it as a "BLARK" initiative, a number of organisations, in particular ELRA, ELSNET, and the LDC (Linguistic Data Consortium, USA), have contributed to the identification, promotion, dissemination, production, etc. of Language Resources and related tools as a support to the HLT community.

Further to the definition of BLARK which focussed on the LRs needed for each language, one may face other existing and more sophisticated tools and systems that are also capable of processing language data. In many cases, sophisticated tools are a combination of many basic tools that require BLARKs to be developed. In many other cases, such sophisticated tools require extended data of their own. A distinction could be then made between several levels of a Language Resource Kit, the first level being a basic language resource kit "BLARK", and the other levels could be referred to as Extended Language Resource Kits or "ELARK".

Further to its own experience and other reports from partners such as the Dutch initiative, ELRA implemented and improved its original matrix which first attempted to cross-link the types of language resources with respect to the languages that could be identified as required languages. In order to understand the needs in a clearer and more complete way, ELRA has extended its matrix to a list of potential applications to be cross-linked with the LRs needed and corresponding languages. The two BLARK matrices as proposed by ELRA aim to be cross-linked, made accessible and modifiable directly from the ELRA web site. This will enable external customers or providers of LRs to fill it in with complementary information and help ELRA at identifying available LRs and promoting the production of new specific ones. At a first step, the combined matrices will be submitted to experts of the HLT field for validation. This could be done through an extended survey and/or the implementation of the matrix online through the ELRA web site. A small part of such matrix is given below:

	Speech Resources	Broadcast speech	Articulatory database	Microp hon e/desk top speech	Read ne wspace r texts	Telephone speech database	Mobile - radio speech	Pronun ciat ion lexicon	On om asti con	Speake r identifi cation spee ch corpus	Lex ica	Monoli ngu al lexicon	Multili ngu al lexicon	Text Co rpora	Broadc ast text corpus	Conve rsation text corpus	News wire text corpus	Monoli ngu al corpus	Multili ngu al and paral lel corpus	Tre ebank
2 Production of documents																				
2.1 Automatic generation (words, sentences, texts)												X	X					X	X	X
2.2 Automatic generation of multimedia documents			X	X				X	X			X	X					X	X	X
2.3 Machine translation												X	X		X	X	X	X	X	X
2.5 Voice translation			X					X	X										X	X
2.6 Speech to speech translation		X	X	X		X		X	X						X					

In a near future, any customer or LR provider aware of an existing LR will be able to complete the cross-linked matrices, pointing to an existing LR. This information will be then considered directly at ELDA in order to check the accuracy of the information. When this information is confirmed, the corresponding cells in the matrix will be filled in accordingly and made available online.

In the future, such an initiative, combined with all ongoing initiatives (and hopefully many more) focussing on the same goal, should contribute to map and, in the end, fill, if not all, at least a fair number of the gaps that should improve the working material of the HLT community. In these initiatives, we should not omit the maintenance work on Language Resources, further to the production work. Indeed, expenses on LRs are big enough to take into consideration their reusability on a long-term, therefore maintenance and updating are rather important and crucial issues.

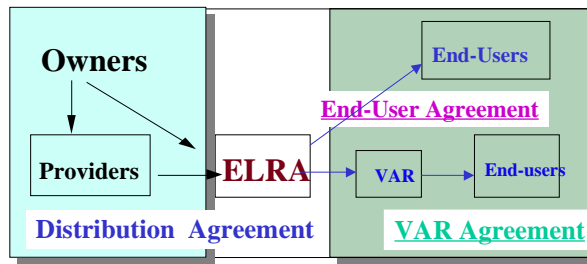
A larger matrix will be discussed at the conference with other languages.

5.2 Handling the legal issues related to the availability of Language Resources

The basic principles of language resource licensing have been worked out with the support of lawyers. At the beginning, marketing Language Resources was a new activity, and creating an equitable and balanced framework was not easy. It was agreed that one of the priority tasks of ELRA was to simplify the relationship between producers/providers and users of LRs.

In order to encourage producers and/or providers of LRs to make such data available to others, ELRA has drafted generic contracts defining the responsibilities and obligations of both parties.

To minimize variations in agreements and to keep things simple, these contracts are based on the following model:



This model takes into account the interest of both parties (producers and users), in keeping with ELRA's role as a neutral, non-profit organization, dedicated to promoting the language engineering field. Contracts (or Licenses) are drawn up between ELRA and the resource provider and/or ELRA and the resource user (either a VAR, Value Added Reseller, or End-user). In this context, it is important to note that ELRA assumes that all resource providers/owners have the appropriate rights to the material that they offer. The generic contracts should be regarded as guidelines for drafting specific agreements. They are based on legal principles, on common sense and market conditions, and are easy to understand, to implement, and to apply. Since 1996, they have evolved in the light of feedback from our members, customers and resource providers.

Let us illustrate the role of ELRA with our resource referenced as ELRA-W0006 which consists of 6 written corpora of similar nature, provided by six different newspapers through Europe (Le Monde from France, Financial Times from UK, Handelsblatt from Germany, Expansion from Spain, IISol24Ore from Italy, and Het Financieele Dagblad from The Netherlands). ELRA has signed one contract with each provider. If this resource is purchased via ELRA, the customer needs to sign one agreement. If the customer rather chooses to go to each individual provider, he/she needs to sign 6 licenses in 6 different judicial systems and will probably have to pay at least 6 different lawyers plus his own!

Contracts between ELRA and LR providers grant distribution licenses by the provider to ELRA. In other words, the purpose of the contract is for the provider to supply the LRs and to receive payment, royalties or other compensation in return. The contract lays down that the LRs must be delivered with any necessary documentation in a specified language. It also stipulates that ELRA is allowed to publicize the existence and availability of the LRs in its catalogues, and to reproduce, and duplicate them for distribution purposes in accordance with its marketing, distribution, and commercialization policy. Alternatively, these tasks can be undertaken by the provider (in this case the terms and conditions must be specified).

ELRA agrees to distribute the LRs and grants its users (i.e. members and customers) the right to use them, in full or in part, for the purposes defined in the agreement between ELRA and the provider, at the user's institution or site.

The contract between ELRA and the user grants the latter a non-exclusive, non transferable right to use, rework and build on the LRs for the purposes agreed upon between the provider and ELRA within the user's institution. To this extent, the user is allowed to create derivative works or software for his/her own internal research and development activities from the LRs or any of their components.

The agreements make no provision for the user to acquire any ownership, rights, title, or interest in the LRs. The user acknowledges the right of the provider in the LRs and related materials including support documentation, and he commit not to infringe them in any way.

One point that has to be clearly defined is what usage is allowed: some providers allow their resources to be used for research and technology/product development, while others only allow distribution for research purposes.

ELRA states that in all cases the user shall not copy or redistribute the LRs, although backup copies may be made. Any use of LRs by an affiliate, subsidiary, or other entity outside the user's place of business must be negotiated with ELRA. In particular, the use of LRs or parts of them in any documentation, application or service which is charged for by the user, may be subject to a separate agreement.

ELRA considers the production and distribution of these licenses as one of its contributions to the development of LR brokerage, so the licenses are available on the Web (as copyrighted documents) but we encourage all actors to use them. One can get electronic copies from the ELRA Web site.

5.3 Distribution activities and Pricing policy

The first two years of activity were devoted to the establishment of the infrastructure and to the identification of valuable resources. This explains the low take of our sales in 1995-1996.

The pricing policy is also a crucial issue that needed careful attention. This had to take into account the fact that we were establishing a new market in which LRs should be traded like any other commodity, as well as the requirements and restrictions imposed by the provider (or the producer) when it comes to the issue of financial compensation. Likewise, market knowledge and contacts with potential providers allow ELRA to always have reliable and useful information on the demands and needs of the

market. The ELRA approach is to simplify the price-setting, to clarify possible uses of LRs, and to reduce the restrictions imposed by the producer.

The prerequisite of acting as a broker is that each purchase renders a payment, covering the compensation claimed by the owner of the resource. In general, ELRA is not the owner of the resources, and can therefore only set a fair price in co-operation with the owner. This co-operation in setting the price is often based on conventional pricing methods like production costs, expected revenues, etc. The pricing must also take into account the ELRA distribution policy, which is to always try to offer a discounted price to its members.

In some cases, the providers accept to have their resources distributed for free. This is often the case when production of LRs is already financed by the European Commission or by national governments. When examining the catalogue, one will notice that the ELRA members benefit from price reductions ranging from 10% up to 70% on the public price. Exceptionally, ELRA is able to offer price reductions even without this being financially supported by the providers. This is just one of the services offered to our members, proving that ELRA is unique in its way of offering services and distributing LRs. The restrictions on the distribution, sometimes imposed by the providers, are more often of two kinds: it is either a restriction on the user profile or a restriction on the usage. The providers may limit the distribution to members only or to Europeans only, or they may restrict the use of their resource to research or even to academic research. When the restrictions are connected with the use, the reason is often that the providers do not want their resource to be used in technical (commercial) development.

The following tables show the situation of the LR distribution via ELRA. These tables highlight the distribution to members vs non-members, to Europeans vs non Europeans, the sales among the different colleges (speech, written, and terminology) and the resources distributed for commercial vs. research use.

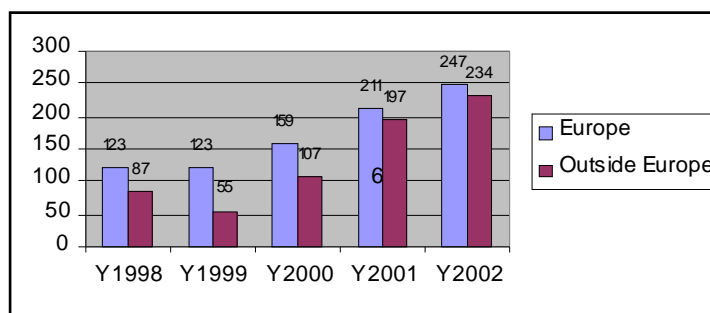
Distribution of LRS to members versus non members:

	1996	1997	1998	1999	2000	2001	2002
Members	11	37	173	107	151	303	303
Non members	6	12	31	71	115	105	178
TOTAL	17	49	204	178	266	408	481

Distribution according to the use of LRs:

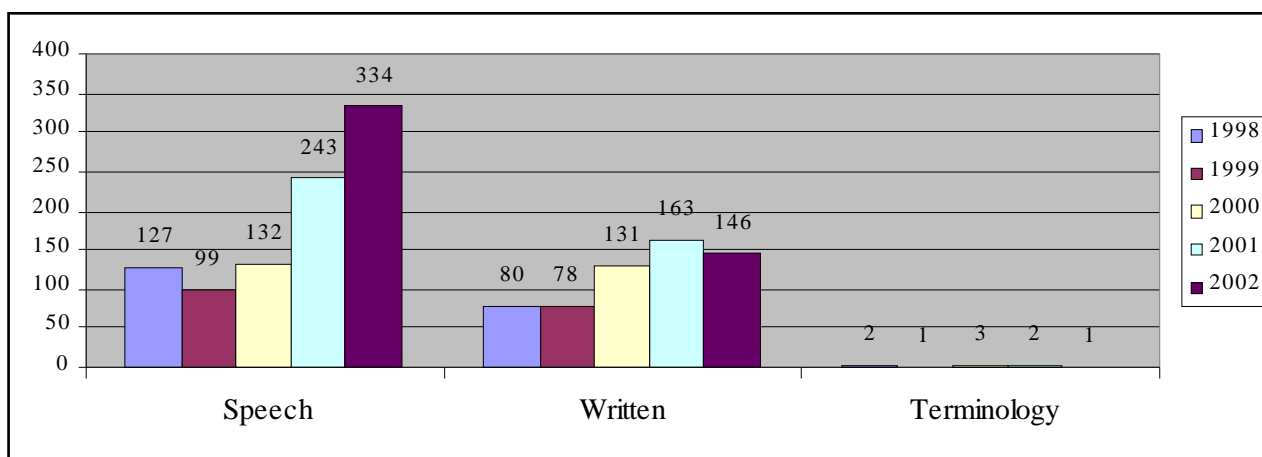
	1998	1999	2000	2001	2002
Research	122	104	156	238	347
Commercial	88	74	110	170	134
TOTAL	210	178	266	408	481

Distribution according to Geographical origin of client (Europe based organisations vs Organizations based outside Europe):



Sales according to the type of LRs over the years (including the Aurora¹ package)

¹ The Aurora project was set up to establish a world wide standard for the feature extraction software which forms the core of the front-end of a DSR (Distributed Speech Recognition) system. ELRA has been asked to be in charge of the distribution of the databases developed for this purpose. Aurora resources are included in the catalogue and whoever wanted to purchase them is welcome to do so under the adequate licensing. The sales of the Aurora databases have more than doubled: 55 databases had been distributed in 2001, 135 in 2002.



Despite our marketing and commercial efforts, we are still making most of our income from spoken language resources followed by written resources. Terminology resources distribution is not taking off so far. A deep analysis is being conducted to understand the rationales behind this.

Most of our customers join ELRA before buying the LRs (which is enforced by our pricing policy). Our contribution to the development of research activities has seen considerable growth and our involvement in research and commercial developments is balanced and shows a substantial increase of the items distributed for R&D. When it comes to the distribution by geographical area, we can see that this development is more significant outside Europe, both in terms of quantity of resources sold and in terms of revenues.

5.4 Validation and Quality assessment

The users of LRs, in particular ELRA's "customers," need to know about the product they are purchasing: they need to know its technical specifications and need to be assured of quality control. ELRA must build up a reputation for the products it sells, though there may be a role for products of limited quality to the extent that they have properties that are of interest to the research community. **Validation** is normally used by ELRA in reference to the activity of checking the suitability for the market, the adherence to standards, and the quality control of the LRs. Of course, "the market" reaction to the product is the ultimate validation.

ELRA therefore had to set up a system to enable a specification and quality control document to be issued with every product that it licenses. A provider that creates a LR and enters into an agreement with ELRA for its distribution will be expected to provide some information about the content of that LR product. ELRA cannot rely only on a specification document since it is the reputation of the Association that is at stake. To promote quality control of its language resources, ELRA installed a Validation Committee (VCom). The VCom addresses a number of critical issues including the following topics which are essential elements of quality maintenance.

- Maintain the LR validation portals
- Collection of specifications for various kinds of LR and making an effort in disseminating best practices and guidelines for LR production.
- Describe the quality of existing LR (Improve the quality of existing LR.) and Validation of LR: relevant checks and procedures
- Encourage user to report bugs through Bug report services for various types of LR via the internet and Ensure that bug reports are exploited to improve the quality of LRs being distributed (e.g. production of Patches for corrected versions of LRs).
- Dissemination of work on LR quality via the web.

For an overview of current activities of the Committee: validation of language resources, standardization, bug reporting, patches of updates of language resources, and dissemination, please refer to <http://www.elda.fr/article14.html>

The definition of the validation methodology required the establishment of co-operations with projects aiming at the production of guidelines, standards, and specifications (e.g. EAGLES, PAROLE, SPEECHDAT, INTERVAL, etc.). Our involvement in validation and quality assessment has seen the release of validation manuals in the area of speech and written resources; these manuals have been made widely available. The validation manuals have been drafted by experts working for a number of institutions that agreed to join a network of validation Units that help ELRA to carry out quality assessment whenever possible and needed.

5.5 Commissioning the production of needed Language Resources & Market watch

ELRA has issued a series of calls for proposals to help sponsor the production, and/or the packaging or customization of existing ones, which are needed by the Language Engineering Community. The purpose of the calls is to ensure that necessary resources are developed in an acceptable framework (in terms of time and legal conditions. Usually each call targets projects with short time scale (projects lasting up to one year) and the size of the funding is very small. ELRA funding is to be seen as effective and useful for producers being both tactical in their aims for the targeted market, which means that they do know all about the needs on the specific market, and strategic with regard to what to produce in order to fulfil these needs. The resources to be selected for funding must be in demand on the market and the resources should preferably be easy to produce, without any technical controversies involved. From its market monitoring, ELRA identifies several key speech and written resources. ELRA then categorizes and prioritizes this set of resources.

For instance within a 1999 Call for proposals ELRA received over 30 proposals that have been screened by a review committee that consisted of the ELRA Board members, a few appointed external experts, and a European Commission (DGXIII - Human Language Technologies sector) representative. The 30 proposals covered all the resources itemized in a preference list disseminated by ELRA. The preference list was derived from our market watch and surveys. ELRA will continue to use its funding capabilities to commission the production of Language Resources. It will also continue to survey the needs of HLT players and report on these to the interested parties (national agencies, European Commission, etc.).

As mentioned above, regular market surveys are conducted by ELRA to monitor the needs of the players in this field. The last one was carried out through questionnaires about the needs, emailed to over 1000 contacts, in addition to all major mailing lists in the field (e.g. Corpora, Linguist, elsnets, etc.). The summaries are given in the ELRA newsletter when appropriate.

5.6 Human Language Technologies Evaluations

This section elaborates on the evaluation activity which was initiated at ELRA progressively since 1998, to support the evaluation of Human Language Technologies. ELRA is exploiting the infrastructure set up for Language Resources business to establish and efficient and cost-effective activity on evaluation: both carrying evaluation campaigns and distributing evaluation packages. The main objective is to play the role of a clearing house with the support of a network of institutions willing to ensure that Europe has its own infrastructure for this crucial domain.

5.6.1 Carrying out Evaluation Campaigns and/or Providing Evaluation packages for all Human Language Technologies.

In order to better serve the Human Language Technologies community, ELRA has decided to set up and manage Evaluation activities in Europe with the support of several partners, capitalizing on the model of Language Resources distribution that proved to be efficient.

Human Language Technologies Evaluation forms a fundamental part of the development of language engineering products. It is essential for validating research hypotheses, for assessing progress and for choosing between research alternatives. It enables R&D teams to assess the impact of innovations on system performance. For example, does changing parameter x entail an increase in system performance validating the change?

Evaluation also identifies promising technology or research directions enabling industry to assess its market value. However, language engineering displays a paradoxical property in that in many areas, the state of the technology has reached a level barely sufficient to be usable in practice. Nevertheless, many commercial language-based applications do exist (e.g. machine translation, text summarisation, dictation, spoken dialogue systems). Comparative evaluation could help clear up the issues, where the advertised performance claims are difficult to assess and compare objectively.

Evaluation also allows funding agencies to determine whether their investment has led to a significant progress. Many national, European or international projects require regular progress reports. Therefore, the results of evaluation campaigns enable the progress of the project to be tracked. It also gives funding agencies the data necessary to quantitatively evaluate the progress made possible by their investment, and thus suggests priorities on where to plan research efforts and support for application development. Evaluation campaigns also provide useful input when deciding whether a technology is mature enough to be considered as a candidate for starting commercial application development.

For ELRA, there is a further side effect of evaluation campaigns is the production of high quality evaluation resources, in the form of training and test data along with evaluation software packages, distributed or produced during evaluation campaigns. Also, the availability of evaluation packages enables all researchers in a particular field to evaluate, benchmark and compare the performance of their systems.

In the USA, the DARPA government funding agency is active in the evaluation of the principal areas of HLT: speech dictation, spoken language understanding, broadcast news transcription, named entities extraction, topic detection and tracking, text retrieval, message understanding, machine translation, speaker verification, character recognition, etc. It organises competitive evaluation campaigns and publishes the results in a workshop. The tasks within the different language technologies have been made more and more difficult, in agreement with the improvement in the various technologies over time. In order to have the necessary logistics for

such evaluations, two entities play a major role in this framework: NIST, the National Institute for Standards and Technology, and the LDC, the Linguistic Data Consortium, which was created for the purpose of distributing language resources.

It would appear that the US-based evaluation programmes follow a top-down strategy i.e. the US government strongly influences the campaigns, but provides abundant funding and a long-lasting infrastructure. In Europe, the strategy has been rather more bottom-up, starting from individual research groups and HLT systems. The US campaigns have inspired efforts at creating a lasting and permanent evaluation infrastructure in Europe. However, the picture in Europe is more fragmented for several reasons. First, there have been much less resources devoted to evaluation and secondly, evaluation efforts have come from many different sources, the result of which is that there is no equivalent European evaluation infrastructure. However, there have been several initiatives, either at the EU level (CLEF, SQALE, TSNLP, the proposed EAGLES evaluation methodology, ETSI/Aurora, DiET, DISC, TEMAA, and SPARKLE etc.), or on a national level (Grace, Aupelf ARC, in France, Verbmobil and the Morpholympics in Germany and SENSEVAL/ROMANSEVAL co-sponsored by several EU-projects, ELSNET, ELRA and the British government). But all these initiatives were funded within limited duration projects, and there is no permanent entity designed to organise evaluation campaigns and capitalise on the resources and packages created during these independent initiatives. Therefore, the result is that European research teams are obliged to evaluate their technologies in US evaluation campaigns, using US evaluation packages which are subject to the geo-political incentives of the US research funding bodies.

From the analysis conducted within a European Funded project called ELSE, comparative technology evaluation (in conjunction with DARPA style competitive evaluation) brings many interesting features. It forces researchers and technology developers to go deeper in their research field when they try to figure out how to measure the performance of a system for a given task. It gives technology developers objective information in order to make choices in system development. It gives industry the possibility of comparing their technology with others by participating in evaluation campaigns, or by acquiring the test data and comparing their systems performance with what has been achieved and reported so far. In particular, it provides SMEs with an efficient and easy market watch.

ELSE has provided recommendations for setting up such an evaluation infrastructure in Europe, which are taken into consideration by ELRA in its development. It has identified the advantages of using the comparative evaluation paradigm and has listed several language technologies which could immediately make use of the evaluation infrastructure based on their relevance for research and industry.

5.6.2A European Evaluation Infrastructure within ELRA

ELDA has a proven track record in the efficient and cost-effective distribution of LRs on both a European and worldwide level. It has set up an organisational model for LR networks dedicated to the specification, commissioning, production, validation, packaging and distribution of LRs with the legal issues resolved.

Along with its experience in national and European evaluation projects, ELDA's evaluation department capitalises on this experience to create an organisational model for efficient and cost-effective evaluation management. This entails the creation of a European, even international, network or infrastructure of evaluation centres providing evaluation resources, software packages, technology, forums of scientific expertise and R&D centres for the independent, ethical evaluation of human language technologies.

The European infrastructure would be organised along two major principles, proactive and reactive evaluation schemes. ELDA's evaluation department is currently taking part in reactive evaluation in that it has been granted national and European projects, such as the French national programmes EVALDA (evaluation of more than 8 technologies ranging from corpus alignment tools to machine translation), AMARYLLIS CLEF (Information retrieval systems) CHIL, TC-STAR_P, etc. ELRA will be involved in the specification and production of evaluation resources, packages and protocols. An exit strategy is defined for each project where the evaluation resources, packages, software and knowledge (final project reports) produced in each evaluation campaign for each linguistic technology is made available to external players through ELDA's catalogue for a modest price. ELDA is well placed to carry out this mission due to its significant experience in the specification, production, packing and distribution of LRs – a related task. For more information on these projects please refer to <http://www.elda.fr/rubrique13.html>.

In parallel, the evaluation infrastructure would be proactive. ELDA endeavours to make available evaluation resources and packages for all linguistic technologies in as many languages as possible. At the very least, this European evaluation infrastructure would have to make available evaluation resources and packages for the official EU languages.

A European initiative is required due to the international and multilingual nature of linguistic technologies. All major developers work on several languages even if they do not create truly multilingual systems. Furthermore, the major players operate on an international level. International cooperation has also been the key to the success of many projects or systems. Therefore, porting linguistic technologies across more and more language barriers leads to a greater need for a multilingual evaluation framework.

Finally, many European language markets are too small to sustain their own evaluation programmes. For example, a language with relatively few speakers i.e. Dutch or Danish, can only rely on European cooperation to organise the evaluation campaign that they need. With the arrival of the new member states in 2004, ELDA faces the challenge of providing evaluation resources and packages

for these new languages and therefore seeks cooperation with the new national agencies, research centres and private concerns to make available, commission and produce language and evaluation resources in the new languages.

It would not have to stop there. ELDA's long term goal in this respect is to cover as many world languages and human language technologies as possible, therefore creating an international evaluation infrastructure, dealing not only with European languages, but languages such as Chinese, Japanese etc.

In either case, the evaluation packages, in the form of training data, test data, test suites, evaluation protocols, software packages, toolkits, agreed methodologies, metrics and even savoir-faire, created through evaluation campaigns or by commissioning in a proactive manner, would be made available to the wider research community via ELDA's catalogue in the same way that ELDA makes LRs available. In this way, ELDA can take on the role of European clearing house or centre for evaluation technology, resources and expertise. In addition, ELDA seeks to standardise evaluation protocols and make these standards available, along with the scientific justification behind it. Furthermore, using its expertise in evaluation, ELDA seeks to advance basic research in the subject of evaluation. In so doing, ELDA would be advancing the field of metrology in language engineering evaluation.

As the middleman of a European evaluation infrastructure, ELDA would also become the forum or focus of knowledge on evaluation issues and evaluation metrology. In the course of evaluation campaigns and the commissioning of evaluation packages, ELDA will have acquired a good deal of expertise in evaluation over the entire range of linguistic technologies. In so doing, ELDA would become a centre of knowledge on evaluation in HLT and would be well placed to disseminate this knowledge.

5.7 Information dissemination, Promotion and Awareness:

5.7.1 Major Conferences established by ELRA: LREC & LangTech

Our contribution to information dissemination activities consisted of the establishment of an important international conference in 1998, the International Conference on Language Resources and Evaluation – LREC. LREC and its satellite workshops were held in Granada, Spain in 1998, in Athens (Greece) in 2000, in Las Palmas (Spain) in 2002, and the next one is planned for May 2004 in Lisbon (Portugal). The first edition attracted over 500 attendees from over 38 different countries and all the continents, the last one in 2002 attracted over 700 representatives from the various HLT fields. The success of the LREC 2002 conference can be illustrated with the following figures: for the main conference, 365 papers were selected out of over 460 which had been submitted and reviewed. The number of papers submitted to the LREC'2004 is over 800 which show the importance of such forum for our field.

LREC is now recognized as the main conference devoted to Language Resources and Evaluation. More information on LREC can be found on the LREC web site, at www.lrec-conf.org.

ELDA has also taken an active role in the establishment of LangTech. LangTech is the European Forum for Speech and Language Technology and covers a wide range of speech and language technologies with three focus areas: speech technologies and applications, semantic web and knowledge management, and applications addressing all issues related to multilinguality.

ELDA was in charge of the organization of LangTEch'2003 which was the 2nd edition. LangTech features relevant developments in many aspects of speech and language technology covering existing speech and language technologies ready for deployment, new solutions ready or close to market, case studies showcasing successful technology transfer, best practice reports on exploitation of such technologies, success stories on the marketing of speech and language technologies, new trends in research and future market opportunities, etc. It also constitutes a forum for meeting with venture capital for companies in the HLT sector. LangTech'2003 was held under the patronage of Mr. Erkki Liikanen, member of the European Commission and with the support of the French Ministries in charge of Research, Industry and Culture, in relation to Technolanguage, the national programme for Language Technology. It took place in Paris on 24th & 25th November 2003. It attracted over 280 participants, whose two-thirds came from the industry, attended the conference sessions and the panels organised at LangTech 2003, which has proved to be a real successful and fruitful forum.

After its 2002 and 2003 editions, it is undeniable that LangTech is the place to be for HLT and related-fields companies, as well as for users of speech and language technologies. More information on LangTech can be found at www.Lang-tech.org.

One of the other means to make ELRA more visible consists of our quarterly newsletter devoted to promoting the filed and announcing new resources available through ELRA. A special issue is usually devoted to LREC with summaries of the opening, closing and several technical sessions. The Web site is another means. We have recorded a larger number of visits to our Web site and the site is updated on a regular basis, with new resource descriptions and documents of interest to the language engineering community, such as validation manuals.

5.7.2 Euromap Language Technologies

Another activity devoted to promoting HLT and contributing to technology transfers and worth mentioning here was implemented within the EuroMap project. Euromap Language Technologies project was a EU-funded project (started in Feb. 2000 for 24 months) aiming to provide awareness, bridge-building and market-enabling services to boost opportunities for market take-up of the

results of national and European HLT RTD projects. The key focus is on accelerating the rate of technology transfer from the research base to the market by creating communities of interest from established and emerging players in the development and value chain.

Project general objectives:

- *increase the number of projects that deliver ready-for-market results;*
- *accelerate awareness of the benefits of HLT enabled systems, services and applications within user sectors, policy makers and national administrations;*
- *boost the number of best-of-class technology developers participating in research projects;*
- *improve the relevance of project targets and technology supplier/user needs;*
- *improve the match between HLT design and supplier/end user expectations;*
- *facilitate user partnerships and communities for beta testing, demonstration, real-time utilization monitoring and other close-to-market application activities.*

Euromap initially focused on the national level, and then extended to cross-border activities, including accession countries. The project published all its results and project-related information on the following web site : <http://www.hltcentral.org/euromap> as well as in a final report.

Project results, achievements and methodology:

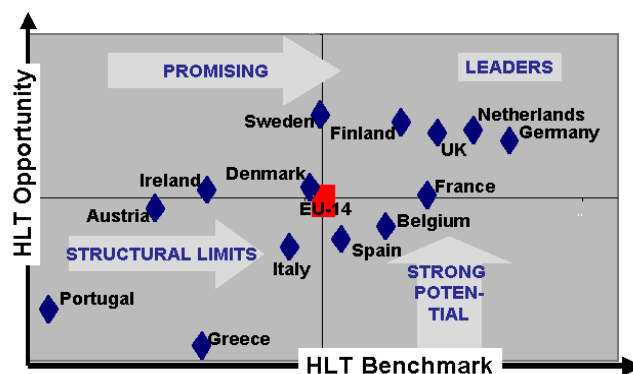
- Web sites (central and national- contain all activities and results contributed to reaching the objectives, News in the field of HLT, Calls for proposals (HLT, eContent, FP, FP6, etc.)
- Newsletter (English and localized versions, monthly electronic newsletter presenting all news related to HLT, more than 3,900 subscribers)
- Knowledge base (Contact database, Fact files for the partner countries, Information about HLT in Baltic, Eastern and Central Europe, and Mediterranean, HLT Policy making Technology in Europe)
- Success stories (stories about successful HLT transfers, written by NFPs and featured cases from the partner countries, 43 success stories)
- Events (organized each year by the NFP partners 30 national seminars and 20 topical seminars)
- LangTech2002: Organization of an international HLT conference with the collaboration of different parties, DFKI, Investitions Bank Berlin, ELDA, VDI/VDE-IT (local organizer), ELSNET, Arax, and Center for Sprogteknologi (CST). The event took place in Berlin, 26-27 September 2002. Result: 330 participants from more than 30 countries, and across 5 continents. Nearly two-thirds of LangTech attendees came from industry or commercial companies.
- Euromap Language Technologies report: Benchmarking HLT Progress in Europe.

The findings of the Euromap Language Technologies project have been synthesized in a final report 'Benchmarking HLT progress in Europe'. The Euromap final report contains an analysis of the market opportunities for HLT in Europe, based on benchmarks of a number of relevant factors. Data from third-party sources and NFP partners' research were used to provide "ratings" of factors that influence the potential for HLT take-up, and findings from Euromap research to assess the maturity of various factors specific to HLT RTD.

The factors measured to assess the relative HLT Opportunity of Member States (and where possible Candidate Countries) include:

- Technology Development and Innovation - including HLT Research Maturity, RTD & Innovation, and HLT Language Politics
- Supply-side Factors - including the environment for new-business formation, access to key market channels, and supply-side "readiness"
- Demand-side Factors - including trade competitiveness, ICT infrastructure and demand-side "readiness".

There were, therefore, three "areas" of maturity (technology development/innovation, supply, demand), and three "benchmark factors" in each area - making a total of nine factors that will be considered in the analysis. These are explained in more detail in the report. For instance the HLT benchmark of European Union countries led to the following diagram which shows the leading countries with important market opportunities for HLT (e.g. Germany, Netherlands). It shows also the ones with strong potentials with mature HLT products but low opportunity now (e.g. Spain). It also highlights the severe structural limits and weak opportunities that are faced by countries like Portugal or Greece.



To have a hard copy of this report please contact: Mahtab Nikkhou at nikkhou@elda.fr; You may have an electronic version of this report from: <http://www.hltcentral.org/euomap> (at the "Reports" section)

5.7.3 NEMLAR: A Network for Euro-Mediterranean LAnguage Resource and human language technology development and support

NEMLAR is another initiative that is useful to the SCALLA conference attendees as it illustrates the need for European organizations and in particular to ELRA to partnership with other organization to boost the Human Language Technologies activities worldwide.

The goal of the NEMLAR (Network for Euro-Mediterranean LAnguage Resources) project, an Accompanying Measure under the INCO-MED programme, is to establish a network of partner centres of best practice in Arabic and other southern Mediterranean language processing dedicated to surveying the state of the art on language resource needs, establishing development priorities, validating the interoperability of components and standards, and developing a minimum set of language resources in order to enable linguistic diversity in the southern and eastern Mediterranean region for digital information and communication on global networks. The consortium includes a consortium of 14 partners from European and Arabic countries. Nemlar started in February 2003 and will last 30 months. The main specific objectives are:

- Qualify and support at least one leading LR actor in each country in the network in partnership with recognised European centres of excellence in Arabic and other indigenous speech and text processing.
- develop and disseminate a 'map' of Euro-Mediterranean stakeholders, national and cross-border projects, and existing language resources and processing tools addressing the existing linguistic diversity in the region (local variants and/or common forms of Arabic and other widely-used local languages)
- survey and analyse the key strengths, weaknesses, opportunities and threats to the development of Arabic and other language resources in the region and establish a set of key priorities for developing LRs
- specify a protocol for a standard set of minimum resources and tools per language (or local form of Arabic according to country) necessary for carrying out all further research and training on language resource development
- develop a standardised training course (with online support) that can be used by Euro-Mediterranean partner experts to train people for priority tasks in creating robust language resources
- try to have an Arabic language track, session or sub-event at major language processing and language resource conferences in Europe in the fields of speech and text
- create a NEMLAR website for collecting and disseminating global information on Arabic and local language resources, tools and technologies, and set hit targets
- raise awareness of the state of play of Arabic and local language resource development among all stakeholders by disseminating a regular information newsletter, ensuring information feeds to existing networks and information sources, and by holding an international conference on current activities and future orientations in Arabic and local language resource creation and management half-way through the project.

Details about the project can be found at <http://www.nemlar.org> in particular a conference will organised in fall 2004 and will be advertised there.

6 ELRA COLLABORATIONS AND PARTNERSHIP

Since its creation, ELRA has been an active partner of COCOSDA and has supported its annual meetings. ELRA also co-operates with other national or regional organizations who are involved in activities relating to those of ELRA. These include LDC (Linguistic Data Consortium), the oriental COCOSDA, AUPELF (association of French speaking academic institutions), etc.

Another crucial collaboration foreseen by ELRA is with the European National Programs. As of today over 7 countries have started their own program in Language Engineering. The expected outcomes include Language Resources. ELRA has been appointed as the distribution channel for many of them.

7 CONCLUSION

Through this paper, we hope to share, with our Asian colleagues, the European Language Resources Association's experience in establishing an infrastructure for the collection and distribution of LRs. It is of paramount importance that regional organizations emerge and co-operate between themselves with respect to the issues described in this paper. The main common task would be to achieve, all together, a better streamlining of efforts in the development of new Language Resources that are of interest to "global" players.